Classification Rule Extraction by Ant-Miner for Weed Risk Assessment

Fukuda, K. [1] and J. Brown [2]

[1] Environmental Science Programme, Department of Mathematics and Statistics, and Department of Computer Science and Software Engineering, University of Canterbury, Christchurch
[2] Department of Mathematics and Statistics, University of Canterbury, Christchurch
Email: k.fukuda@math.canterbury.ac.nz

Keywords: *Ant-Miner, Classification, Risk, Weeds*

EXTENDED ABSTRACT

Weed risk assessment (WRA) models developed by Pheloung et al. (1999) and Daehler et al. (2004) allow an informed decision prior to introducing potentially invasive plant species into a country. In this study, Ant-Miner, a data mining tool, is used to develop classification rules for WRA models of Australia, and Hawaii and the Pacific.

Ant-Miner (Parpinelli et al., 2002), based on Ant Colony Optimization (ACO), is a metaheuristic inspired by the foraging behaviour of ant colonies. Its objective is to solve discrete optimisation problems and extract classification rules by simulating the behaviours of ants. For this study, Ant-Miner identifies a shortest pathway described by nodes, i.e., the 50 questions from WRA, by overcoming ant behaviour problems, e.g., the dead end, loop, returning root and evaporation of pheromones (Figure 1), during the search for the destination, e.g., a single decision described by either yes, no or blank to classify the class: low to high risk (*reject*) or *evaluate* and *more information required* in the WRA models. The purposes of detecting the dominant pathway are: 1) to understand how the decision process for plant risk is assessed from answering the questions in the current WRA model, and 2) to understand the WRA criteria in regards to how the decision process differs among regions and climates, e.g., Australia, and Hawaii and the Pacific.

Ant-Miner is found to be an effective alternative data mining tool, since it obtained reasonably high classification accuracy (via 10-fold cross validation); in particular for the Hawaii and Pacific Island WRA model (81±1.24%) and for the Australia WRA model (71±2.26%). The extracted rules for Ant-Miner suggest that high risk species are assessed mostly under the following key factors: for Australia, if the species have been naturalized beyond their native range and reproduce by vegetative propagation, and for the Pacific, if the species have been naturalized beyond their native range and are congeneric, but not parasitic. Ant-Miner detects that the dispersal mechanism is an important factor for the classes *low* or *evaluate* for both Australia, and Hawaii and the Pacific WRA models. On the other hand, from both WRA models, the question about the plant type was found to be less significant for the plant risk assessment. The reproduction process for Australia and the location of the weed for Hawaii and the Pacific are detected to be overall important factors for the plant risk assessment.

Identifying influential factors in weed risk helps improve cost effective biosecurity assessment by highlighting important and modifying or perhaps removing unimportant questions of the current WRA model to increase the overall accuracy. This study will encourage further investigation with larger data sets from different regions in future to add knowledge to help the WRA model improvement.

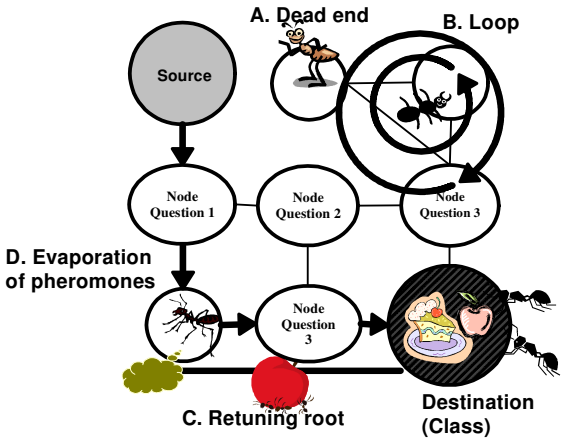


Figure 1. Diagram of ants building a solution.

1. INTRODUCTION

Effective strategies to mitigate and control existing or future invasive organisms are important for maintaining and protecting our healthy ecosystem. Entering and spreading invasive weeds (alien plants) can threaten the native environment, as they can alter the fundamental structure of the ecosystem by changing its composition, structure, and function (Yeates and Williams, 2001).

The weed risk assessment (WRA) model (Pheloung et al., 1999) provides an informed decision prior to introducing potentially invasive plant species into the country. The WRA is established as a biosecurity tool to evaluate new plant introduction in Australia, and as been tested and modified to adapt to the unique climate and environment of different countries, for example, New Zealand (Pheloung et al., 1999) and Hawaii and the Pacific Islands (Daehler et al. 2004), referred to here as Hawaii/Pacific. The WRA models have 50 questions about the main attributes and impacts of weeds to allow assessment of their weediness (see the blank WRA sheet in Pheloung et al., 1999). Individual plant species are assessed by answering questions in the WRA model, resulting in a score from -14 (benign taxa) to 29 (maximum weediness). The total score is then evaluated into three possible recommendations: accept the plan for import (score < 1), further evaluation required for the plant (score from 1-6), and reject the plant for import (score > 6). Additionally, in the WRA for Hawaii/Pacific, a second screening process is applied for scores from 1-6 to determine a further recommendation to either accept or reject (see detailed criteria in Daehler et al., 2004). Daehler et al. (2004) found from a comparison between the WRA and experts' opinions, the second screening process for the WRA improves the number of correctly identified non-pests, i.e., non-pest classification accuracy with the second screening is improved to 85% from 66% without, as well as classifying additional minor pests as non-pests.

Use of the WRA model as a decision making tool is beneficial, since it eases the border security process of plant risk assessment. However, some key issues are of concern to set up such a model. For example, the WRA process is not part of the legal process to prevent importing unless the plant is stated in the State or Federal Noxious Weed List (Daehler et al., 2004). Minimising biases is important as personal opinion on assessing *invasiveness* of weeds can vary among different fields of expertise (Pheloung et al., 1999). It is important to produce a model that describes the phenomena more accurately; this can perhaps be achieved by understanding and increasing knowledge about the model itself.

In this study, Ant-Miner, a data mining tool, based on the Ant Colony Optimization (ACO) algorithm, is used to develop classification rules for WRA systems for Australia and Hawaii/Pacific. This study identifies the shortest pathway to classify each plant species (as either *accept*, *evaluate* or *reject*). The purpose of discovering such knowledge is to help plan the time and cost effective future WRA by identifying important or unimportant questions. For example, if a particular question is found to be important for judging high-risk plants, then this question may be highlighted as important to answer. If it is impossible to answer because the species is new or there is a lack of resources for the new environment, then this question may be divided into a few specific detailed questions. On the other hand, questions that are found to be less important can be removed from the WRA systems. At the same time, if the question is too difficult to answer, then the plant is classified as *evaluate* or *more information required* (as answers tend to remain blank). If some particular questions are more likely to be unanswered, it would be best to identify these and narrow or even remove the types of question that cannot be easily answered. In fact, the studied data sets from Australia and Hawaii/Pacific contained less than 20% and 10% respectively of *evaluate* or *more information required* responses. Hence, understanding about the model may further increase classification accuracy and improve the WRA process.

Ant-Miner (Parpinelli et al., 2002), developed based on Ant Colony Optimisation (ACO), is a metaheuristic inspired by the foraging behaviour of ant colonies, i.e., tracking of pheromones, with the objective of solving discrete optimisation problems, developed in 1980s by Dorigo and Stützle (2004). Due to its nature, ACO has been applied to the travelling salesman problem, and various other fields (sequential ordering, flow shop scheduling and the graph coloring problem (details in Dorigo and Stüzle, 2004), though its application in environmental science is still uncommon.

This paper briefly describes the ACO algorithm, then introduces the Ant-Miner algorithm. In this study, Ant-Miner software (Parpinelli et al., 2002) is used with a slight modification. Generally, Ant-Miner produces N solutions or paths with an overall classification accuracy for N-fold cross validation. In this study, classification accuracy is obtained from N-fold cross validation, e.g., N=10, and also, to allow an interpretation of the path, a single solution (path) is obtained for the whole

data set. Results are focused on major findings, observed from detecting the shortest pathway for the different plant risks, and will discuss how the WRA systems for Australia and Hawaii/Pacific consider the questions differently. Conclusions will include how knowledge discovered via a data mining tool helps plan the cost and time effective WRA model for the future.

2. METHODS

2.1. Ant Colony Optimization

The Ant Colony Optimization (ACO) algorithm is swarm intelligence that is generated by mimicking real ant behaviour. Ants write, read and estimate the amount of pheromone trail (proportional to the utility of using a particular arc) to build a good solution (Dorigo and Stützle, 2004). The stronger the pheromone trail, the higher its desirability. Ants follow a probabilistic decision biased by the amount of pheromone. If no pheromone trail exists, ants move randomly (García-Martínez and Herrera, 2007). A brief explanation of the Simple ACO (S-ACO) algorithm follows.

Let $G = (N, A)$ be the graph to each arc (i, j), and an associated variable τ_{ij}, the pheromone trail. Assume all the arcs A have a constant amount of pheromone ($\tau_{ij} = 1$, $\forall (i, j) \in A$) at first. Then, a probability P is defined for an ant k travelling from a node i to the next node j using τ_{ij} as follows,

$$P_{ij}^{k} = \begin{cases} \dfrac{\tau_{ij}^{\alpha}}{\sum_{l \in N_i^k} \tau_{ij}^{\alpha}}, & \text{if } j \in N_i^k \\ 0, & \text{if } j \notin N_i^k \end{cases}, \qquad (1)$$

where α ($\in \{s, l\}$ when s and l are short and long branches respectively) is a parameter defining the relative importance weight of the pheromone trail, and N_l^k is the neighbourhood of ant k in node i that contains all the nodes directly connected to node i in the graph $G = (N, A)$, but excludes the predecessor of node i (the last node that the ant visited before moving to i) so as to avoid the ants returning to the node they visited immediately before node i. When N_l^k is empty (a dead end, an example is seen in Figure 1-A), node i's predecessor is included into N_l^k. During this process, ants receive pheromone several times by going back and forth; consequently, this can lead to loops (seen in Figure 1-B). Loop elimination is carried out by an iterative scanning process; the path from the destination node back to a given node is scanned. If another instance of the node is reached along the way, the subpath from this instance back to the original instance of the node is a loop, which can be eliminated.

Let a change of amount of pheromone be $\Delta\tau^k$, deposited by the k^{th} ant on arc (i, j) that is visited during their return travel (Figure 1-C),

$$\tau_{ij} \leftarrow \tau_{ij} + \Delta\tau^{k}. \qquad (2)$$

When an ant deposits pheromone earlier than one travelling a longer path, it deposits more pheromone on the shorter path. At the same time as updating the pheromone trail, pheromone trail evaporation (Figure 1-D) is considered, to avoid all ants moving toward a suboptimal path by converging; losing pheromone intensity favours the exploration of different paths. Let ρ be a parameter, where $\rho \in (0, 1]$, then when ant k moves between nodes, the pheromone trails are evaporated as

$$\tau_{ij} \leftarrow (1 - \rho)\tau_{ij}, \qquad \forall (i, j) \in A. \qquad (3)$$

A complete cycle of an iteration of ACO involves pheromone evaporation and deposition, and ant movement.

2.2. Ant-Miner

The following section briefly introduces the main theoretical modifications of Ant-Miner from the ACO algorithm. Ant-Miner is similar to the decision tree algorithm, such as C4.5 (Quinlan, 1993) that discovers the classification rules by following a divide-and-conquer approach:

IF < term1 and term2 and ...> THEN <class>

However, the heuristic functions for decision tree algorithms and Ant-Miner differ in how they consider the entropy; for the former they are computed for an attribute as a whole, but the latter computes them for an attribute-value pair only (Parpinelli et al. 2002).

The procedure of discovering classification rules is as follows. Firstly, an ant starts with an empty rule and adds one term at a time to its current partial rule until one of the two following conditions are satisfied:

1) Adding any term to the rule would result in it covering less than a user-specified minimum number of cases.

Table 1. The original proportion of the class and classification accuracy using the Ant-Miner for the WRA models for Australia, and Hawaii and Pacific.

Class	**Reject**	**Accept**	**Evaluate/More information**
Australia	131 (80%)	3 (1%)	20 (13%) for evaluate 9 (6%) for more information
Class	**High risk**	**Low risk**	**Evaluate**
Hawaii and Pacific	176 (32%)	321 (58%)	58 (10%)
Ant-Miner	**Accuracy rate on test set**	**Rules number**	**Conditions number**
Australia	71.02% +/- 2.26%	6.3 +/- 0.15	13.9 +/- 0.62
Hawaii and Pacific	80.15% +/- 1.24%	7.6 +/- 0.16	21.7 +/- 1.58

2) All attributes have already been used by the ant to create the rule antecedent.

Secondly, the rule can be pruned to eliminate irrelevant terms and thirdly, the amount of the pheromone is increased in the trail followed by the ant and decreased elsewhere (evaporation). Then, newly updated pheromone guides other ants to construct the rule until one of the following is satisfied:

1) Number of constructed rules is equal to or greater than the user-specified number of ants.

2) When the exact same rule has been created by a user-specified number of successive ants.

Detailed algorithms are described in Parpinelli et al. (2002). To operate a data mining algorithm, Ant-Miner modifies the P_{ij} function (originally equation 1 from ACO) which allows the current ant to iteratively add one term at a time to its current partial rule. Let η_{ij} be a value of the heuristic function to estimate the quality or precise value of the entropy associated with the arc (i, j) to improve the predictive accuracy of the rule in Equation 4, where I is the total number of attributes, J_i is the number of values in the domain of the i^{th} attributes and x_i is set to 1 if the attribute Ai was not yet used by the current ant or to 0, otherwise.

$$P_{ij}^{k} = \frac{\eta_{ij}\tau_{ij(t)}}{\sum_{i=1}^{I} x_i \sum_{j=1}^{Ji} (\eta_{ij}\tau_{ij(t)})} \quad (4)$$

Pheromone updating (equation 2 for the ACO) is calculated from:

$$\tau_{ij}(t=0) = \frac{1}{\sum_{i=1}^{I} Ji}, \quad (5)$$

which is inversely proportional to the number of values of all attributes. Then, the pheromone update can be carried out by increasing and decreasing for arcs that are used or not used, respectively (details in Parpinelli et al. (2002) and the equation 3 from the ACO). Ant-Miner parameters are defined by the experiments of running a few different parameter settings, and the best results, e.g., higher classification accuracy, are introduced in the following sections. Note that all classification rules are pruned.

2.3. Data set and Ant-Miner

The data set is taken from the website of the Institute of Pacific Islands Forestry, Pacific Island Ecosystems at Risk (PIER, 2007); http://www.hear.org. The data source shows two types of risk assessments on WRA models; risk assessments for species that are listed on PIER, and not listed on PIER. Both sets of data have the score for a single plant species that is assessed by the Australia and Hawaii/Pacific WRA models; 163 and 555 plants are assessed by the Australia and Hawaii/Pacific models respectively.

The original WRA questionnaire blank sheets are not described in this paper, but are accessible from Pheloung et al. (1999) for Australia and Daehler et al. (2004) for Hawaii/Pacific, or the data source website. Both WRA models have 8 sections and are divided into several questions, and a total of 50 questions. Some questions, e.g., 4.10 from WRA, are different between the two models, as the Hawaii/Pacific model was adjusted from the Australian model.

As previously discussed in the introduction section, the total score for each plant is categorised as a class. The Australia model has four classes; reject (score > 6), evaluate (1 to 6), evaluation or more information (score > 4, but majority of questions unanswered) and accept (< 1). The Hawaii/Pacific model has three classes; high risk (> 6), low evaluate (1-6) and accept (< 1). Note

that the Hawaii/Pacific model has a second screening process for the class *evaluate*, but the second screening process is not used and classified as all *evaluate*. Table 1 shows the proportion of each class. The Australian model classified most of the plant species as reject (80%) compared with accept (3%), but the Hawaii/Pacific model classified most as low risk (58%), followed by high risk (32%). Both models contained about 10% of the plant species that requires further evaluation or more information.

2.4. Ant-Miner program

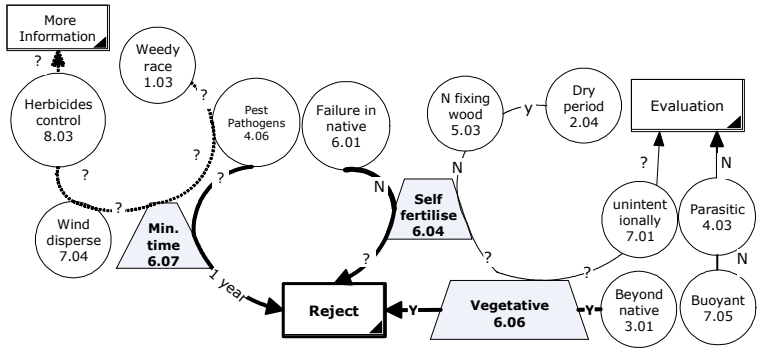
The Ant-Miner program, developed by Parpinelli et al. (2004) detects classification rule using 10-fold cross validation, which divides the data set into ten mutually exclusive partitions, with nine partitions used to extract the rule and the rest used to test the rule, providing the classification accuracy. In this study, the Australia and Hawaii/Pacific WRA models are analysed separately using 10-fold cross validation with four different parameter settings; three parameters (*min cases per rule* = 10, *max uncovered cases* = 10, and *no rules converg* = 10) are kept the same, but the number of ants was changed to 50 and 100, and applied to two numbers of iterations, 25 and 100 respectively. The parameter settings that provided the best-represented results, i.e., the highest classification accuracy, are used to obtain the classification accuracy.

While the original Ant-Miner programme (Parpinelli et al. 2004) was only the 10-fold cross validation method that provides individual classification rules for each of the 10 partitions, in this study, the programme was modified to provide a single classification rule, based on the entire data set. This classification rule was then used to understand the structure of the shortest pathway.

3. RESULTS AND DISCUSSIONS

Table 1 shows the classification accuracy obtained from the best parameter setting; the number of ants

A. Australia WRA model.



B. Hawaii and Pacific Islands WRA model.

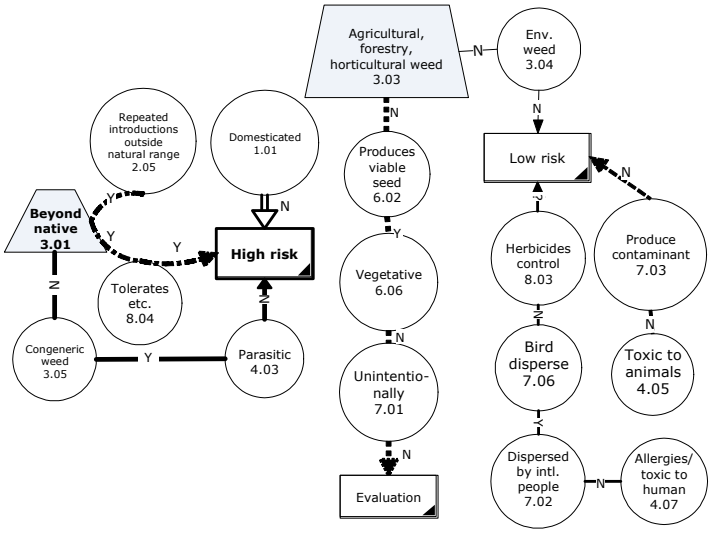


Figure 2. The classification systems for the Australia, and Hawaii/Pacific WRA models obtained by Ant-Miner. Note that the arrows are one way and line styles indicate pathways to classes.

and iterations are 100. Ant-Miner successfully obtained reasonably high classification accuracy, in particularly, the Hawaii/Pacific model data set is found to be more suitable for Ant-Miner, as the higher classification accuracy is detected for the Hawaii/Pacific model (80.15%) than the Australia model (71%).

Figure 2 shows for both Australia and Hawaii/Pacific the best classification rule demonstration, identifying the shortest pathway or important attributes (questions) to lead the different plant risks.

3.1. Australia WRA system

Figure 2-A shows that three common questions were detected among classes; if reproduction by self-fertilisation (6.04) is unknown, this connects to the classes of *reject* and *evaluation*, if the vegetative propagation reproduction (6.06) is true (*yes*) and unknown, this connects to *reject* and *evaluation* respectively, and if the minimum generative time for reproduction is one year (6.07) and unknown, this connects *reject* and more information required, respectively. Besides the above, three pathways were detected for the high risk plant (reject), when the plant is beyond native (3.01), there is no evidence of substantial reproductive failure in the native habitat (6.01) and unknown host for recognised pests and pathogens (4.06).

The pathways for *more information required* were created by all questions – weedy race (1.03), minimum time (6.07), wind disperse (7.04) and herbicide control (8.03) – which are all unanswered (indicated by a question mark in Figure 2); this is a reasonable finding, as more unanswered questions lead to requiring more information about the plant. This may suggests that these questions may need to be improved by adding more specific questions to help in answering them. If these questions are in fact difficult to answer, perhaps even removing them may help the overall analysis, though note that it is important to keep the question about the minimum reproduction time (6.07) because it was found to be important for judging the class.

Interestingly, a common decision making process for all classes was detected to involve the reproduction questions (section 6 in the WRA). This suggests that improving the reproduction question for the plant species by setting up more specific and detailed questions may increase sensitivity and help overall judgement. On the other hand, questions identified as related to the class of *more information* may be removed or have aspects changed to ease answering further, which may help creating the cost and time consuming WRA analysis for the Australian WRA system.

3.2. Hawaii/Pacific WRA system.

Figure 2-B shows independent structures that the questions do not overlap between *reject* (left side of Figure 2-B) and *low risk* and *evaluation* (right side of Figure 2-B). This suggests that the Hawaii WRA system has a strong structure to make a decision for high risk plants, which are assessed particularly (as used twice to form two pathways) by whether the plant is beyond native or not (3.01). If the plant is introduced outside its native range (2.05), and is beyond native (3.01) and tolerates or benefits from mutilation, cultivation or fire then the plant species is rejected. However, if the plant is not beyond native, but is recognised as congeneric weed (3.05) and parasitic (4.03), then the plant species is rejected. Also, if the plant is not domesticated (1.01), then the plan species is rejected. The *low risk* and *evaluation* classes are commonly assessed, when the weed is not found from agriculture, horticulture or forestry (3.03).

3.3. Assessment trends of the WRA between different regions and climate.

The Ant-Miner classification summary (Figure 3) shows that question 5, plant type, was not selected to construct any shortest decision making pathway for both the Australia and Hawaii/Pacific WRA

		Section numbers from the WRA model		
The WRA questions	**Australia**		**Hawaii and Pacific Islands**	
Domestication/cultivation	**1**	03	**1**	**01**
Climate and distribution	**2**	04	**2**	**05**
Weed elsewhere	**3**	**01**	**3**	**01**, **01**, 03, 04, **05**
Undesirable traits	**4**	03, **06**	**4**	**03**, 05, 07
Plant type	**5**		**5**	
Reproduction	**6**	**01**, 03, 04, **04**, 06, **06**, 07, **07**	**6**	06
Dispersal mechanisms	**7**	01, 04, 05	**7**	01, 02, 03, 06
Persistence attributes	**8**	03	**8**	03, **04**

Figure 3. The key WRA questions followed by a section number detected by Ant-Miner as nodes. Numbers in bold indicate classification for reject or high risk plant species.

systems to classify for plant risks. It suggests that the plant type may not be particularly significant for assessing the plan risk. A significant difference in selecting important factors was detected between the Australia and Hawaii/Pacific WRA systems. The Australian system tends to consider the reproduction of the plant species (question 6) as the most important factor, but this was not important for the Hawaii/Pacific system. The Hawaii/Pacific system, on the other hand, selects the place of weeds (question 3), undesirable traits (question 3) and the mechanisms of dispersal (question 7) as important for judging the plant species in regard to their risks. While the high plant risk classification pathway (shown in bold in Figure 3) tends to be assessed by weed reproduction method for Australia and weed location for the Hawaii/Pacific model, the classifications lower than high risk plant (non bold in Figure 3) such as evaluation, low risk or more information required tend to be assessed commonly by the weed dispersal mechanisms (question 7).

This investigation suggests that the fundamental structures of the WRA systems between different climates and regions differ. In order to improve the WRA, reproduction process, reproduction and dispersal mechanisms and weed location are specifically important questions. If these questions can be more specific and allow the assessment to be more accurate, the overall classification may be improved.

4. CONCLUSION

Ant-Miner data mining tool successfully identified the shortest pathway that is the most dominant and important pathway, to classify different plant risks. Examining different WRA systems provides ideas on how regions with different climates have different risk. Generally, for assessing the high risk plant species, the Australian and Hawaii/Pacific systems selected the reproduction process and the location of the weed as important factors respectively. For evaluation or low risk plant species, both models selected the dispersal mechanisms are important. Neither model selected the plant type as an important factor for assessing the plant risks. This may suggest that this question may require modification to be more specific or even may be removed, because it did not help the assessment as compared with other questions.

Identifying influential factors from the model helps construction of cost effective biosecurity strategies. It can target which questions are required to be more specific in order to help construct accurate models. This study shows Ant-Miner can be a useful data mining tool, as it successfully provided important pathways for assessing different risks. At this stage, this investigation was not for constructing new risk-models, instead it was to increase knowledge about the existing model. In the future, many more different plant species and data points taken from different regions will be investigated to help improve the WRA model.

5. REFERENCES

Daehler, C. C., J. S. Denslow, S. Ansari and H. Kuo. 2004. A risk assessment system for screening out invasive pest plants from Hawai'i and other Pacific Islands. *Conservation Biology* 18:360-368.

Dorigo, M. and T. Stützle (2004), Ant Colony Optimization, MIT Press/Bradford Books, Cambridge, MA.

García-Martínez, C. and F. Herrera (2007), A taxonomy and an empirical analysis of multiple objective and colony optimization algorithms for the bi-criteria TSP, *European Journal of Operational Research*, 180, 116-148.

Parpinelli, R.S., H.S. Lopes and A. A. Freitas (2002), Data Mining With and Ant Colony Optimization Algorithm, *IEEE Transactions on Evolutionary Computing*, 6 (4), 321-332.

Pheloung, P.C., P.A. Williams and S.R. Halloy (1999), A weed risk assessment model for use as a biosecurity tool evaluating plant introductions, *Journal of Environmental Management*, 57, 239-251.

PIER (2007), Institute of Pacific Islands Forestry Pacific Island Ecosystems at Risk (PIER) Plant threats to Pacific ecosystems, http://www.hear.org/pier.

Quinlan, J.R. (1993), C4.5: Programs for Machine Learning. San Mateo, CA, Morgan Kaufmann.

Yeates, G.W. and P. A. Williams (2001), Influence of three invasive weeds and site factors on soil microfauna in New Zealand, *Pedobiologia*, 45, 367–383.